A Gazetteer and Georeferencing for Historical English Documents

A Gazetteer and Georeferencing for Historical English Documents

Claire Grover
School of Informatics
University of Edinburgh
grover@inf.ed.ac.uk

Richard Tobin
School of Informatics
University of Edinburgh
richard@inf.ed.ac.uk

Abstract

We report on a newly available gazetteer of historical English place-names and describe how it was created from a recent digitisation of the Survey of English Place-Names, published by the English Place-Name Society (EPNS). The gazetteer resource is accessible via a number of routes, not currently as linked data but in formats that do provide connections between a number of different datasets. In particular, connections between the historical gazetteer and the Unlock[1] and GeoNames[2] gazetteer services have been established along with links to the Key to English Place-Names database[3]. The gazetteer is available via the Unlock API and in the final part of the paper we describe how the Edinburgh Geoparser, which forms the basis of Unlock Text, has been adapted to allow users to georeference historical texts.

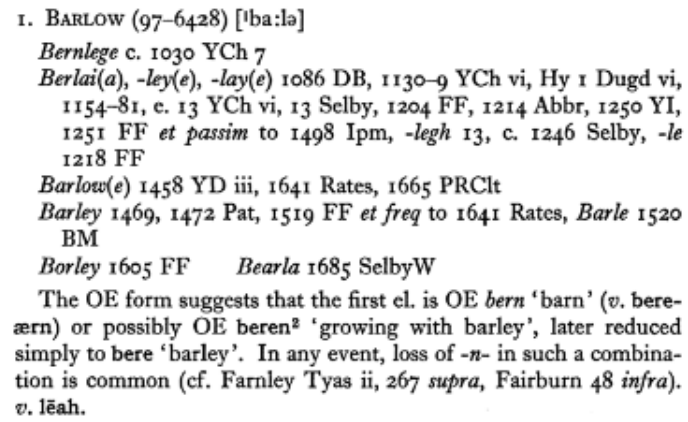
1 Introduction

Place and time are important concepts for historical scholarship and it has frequently been observed that an ability to examine document sets through spatial and temporal filters is one that is highly useful to historians. Georeferencing (or geoparsing) is therefore a technology that has been applied to historical data in numerous projects (for example Hitchcock et al. (2011), Crane (2004), Rupp et al. (2013), Isaksen et al. (2011), Hinrichs et al. (to appear 2014), Grover et al. (2010)). A significant problem, however, is that available georeferencing tools are mostly only able to access modern gazetteer information, meaning that place-names that have changed over time are less likely to be recognised and are highly unlikely to be properly grounded to correct coordinates. The problems can be illustrated with two examples, first the name *Jorvik* which is a well-known historical name for the English city of York and second the name *Bearla*, a historical name attested in a document from 1685 for the modern settlement Barlow in the West Riding of Yorkshire (now North Yorkshire). A first observation is that a named entity recognition (NER) component of a georeferencing system may or may not recognise these as place-names: recognition will depend on the lexical resources or training data used as well as the document context in which the name occurs. Assuming both the names can be recognised, they must then be disambiguated with reference to a gazetteer so that coordinates can be assigned to them. A search for the names using Unlock Places, which provides access to Ordnance Survey records, returns no results for either. A search in GeoNames returns York for *Jorvik* but nothing for *Bearla*. Another historical form for Barlow is *Borley*: both GeoNames and Unlock have records for a modern Borley in Essex but this is clearly not the correct interpretation of the historical *Borley*. These examples illustrate some of the problems and indicate that a historian wanting to georeference a particular document will get patchy output at best from current technology.

The Digital Exposure of English Place-names (DEEP) project[4] has addressed these issues by digitising and processing the Survey of English Place-Names to create the DEEP Historical Gazetteer (DHG). Below we first describe the Survey of English Place-Names and then explain how we have used XML-based language processing tools to convert the digitised volumes into the DHG and other structured resources. We outline the ways in which the resources are made available to users and we discuss the modifications we have made to the Edinburgh Geoparser to allow users to

[1] http://edina.ac.uk/unlock/
[2] http://www.geonames.org
[3] http://kepn.nottingham.ac.uk/
[4] http://englishplacenames.cerch.kcl.ac.uk/

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 119–127,
Gothenburg, Sweden, April 26 2014.



1. BARLOW (97–6428) [ˈba:lə]

Bernlege c. 1030 YCh 7

Berlai(a), *-ley(e)*, *-lay(e)* 1086 DB, 1130–9 YCh vi, Hy 1 Dugd vi, 1154–81, c. 13 YCh vi, 13 Selby, 1204 FF, 1214 Abbr, 1250 YI, 1251 FF *et passim* to 1498 Ipm, *-legh* 13, c. 1246 Selby, *-le* 1218 FF

Barlow(e) 1458 YD iii, 1641 Rates, 1665 PRClt

Barley 1469, 1472 Pat, 1519 FF *et freq* to 1641 Rates, *Barle* 1520 BM

Borley 1605 FF *Bearla* 1685 SelbyW

The OE form suggests that the first el. is OE *bern* 'barn' (*v.* bereærn) or possibly OE beren[2] 'growing with barley', later reduced simply to bere 'barley'. In any event, loss of *-n-* in such a combination is common (cf. Farnley Tyas ii, 267 *supra*, Fairburn 48 *infra*). *v.* lēah.

Figure 1: Survey entry for Barlow in the West Riding of Yorkshire

georeference their historical documents. We conclude by discussing some outstanding issues and consider the steps that will be needed to turn our resources into linked data.

2 The Survey of English Place-Names

The Survey of English Place-Names is a scholarly body of work aimed primarily at readers interested in the origins and development of the place-names of England. The Survey is arranged by historic counties, with the first volume, from 1925, covering Buckinghamshire, and the most recent volume, published in 2012, dealing with part of Shropshire. In the early volumes the Survey was largely limited to major place-names, i.e. the names of towns and villages, but from the 1950s onwards the volumes have become more complex and include many minor names as well as field-names and street-names. More recently treated counties are described across multiple volumes and the growing scale of coverage has meant that there are still some counties which are only partly covered or not covered at all. Nevertheless, the vast majority of the English counties have been surveyed and the resulting body of work is a valuable resource for scholars of many kinds.

Figure 1 shows an excerpt from p. 23 of Vol. 33 of the Survey (1961) which covers the Wapentakes of Barkston Ash, Skyrack and Ainsty in the West Riding of Yorkshire. The excerpt shows the start of the entry for Barlow, the first settlement described in the parish of Brayton. In brackets after the name is an Ordnance Survey (OS) map reference followed by an indication of the pronunciation of the name. Next comes a block of historical forms of the name with information about their attestations. For example, *Bernlega* is attested in a document dated around 1030 referred to by the abbreviation YCh 7, standing for volume 7 of Early Yorkshire Charters (ed. W. Farrer, C. T. Clay, 1914-55). A set of related forms, *Berlai*, *Berlaia*, *Berley*, *Berleye*, *Berlay*, and *Berlaye*, have been attested in several documents ranging from the Domesday Book (DB) in 1086 through to Inquisitions post mortem (Ipm) in 1498. Other forms, including *Borley* and *Bearla* discussed above, follow. The final paragraph deals with the etymology of the name, relating it to the Old English words for 'barn' or 'barley' (and for the second part to the element lēah meaning 'clearing').

3 Conversion to structured format

It can be seen from Figure 1 that the Survey provides a wealth of information with the potential to be useful for many purposes and, in particular, it contains precisely the kind of information that is needed to make a historical gazetteer. In the DEEP project we have digitised all 86 volumes of the Survey and have processed the output of OCR to convert it into a structured format.

As the example illustrates, the format of the volumes is semi-structured with a fairly consistent use of style and font to indicate various kinds of information (e.g. bold font for etymological elements, italics for historical forms). The layout is extremely important with every comma and tab contributing to the interpretation of the information. For this reason, OCR quality needs to be exceptionally high and we have been fortunate that our digitisation partner was able to ensure this high quality. As the survey was created over a period of decades under the supervision of several editors, there is some variability in format across the volumes. The most pertinent variation concerns grid references: early volumes either do not have any or use grid references that cannot be converted to

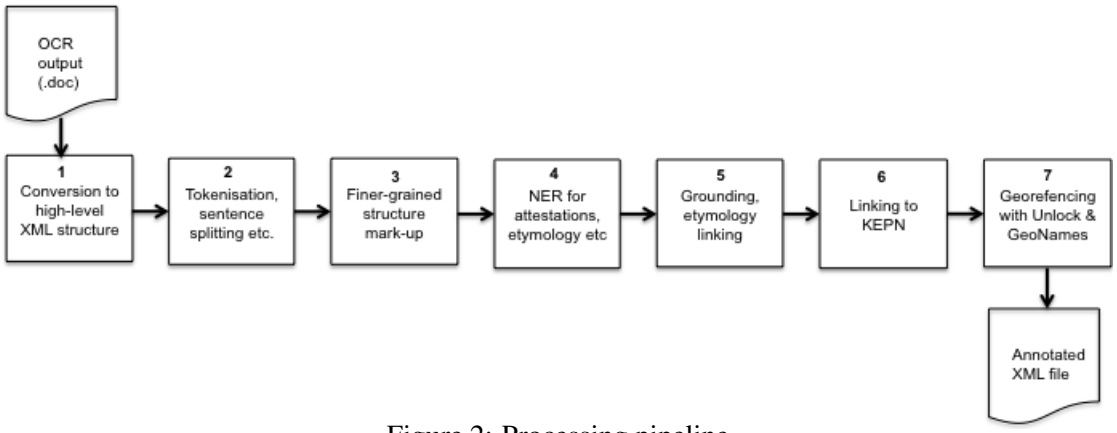


Figure 2: Processing pipeline

latitude/longitude. In many of the later volumes modern OS grid references or older OS sheet-number grid references are provided for main settlements within parishes and sometimes for more minor settlements. However, many of the later volumes do not contain any grid references at all and a significant part of the processing deals with these problems (see Section 3.2).

3.1 Processing pipeline

Figure 2 shows the XML-based processing pipeline that we have created for converting from the OCR output of a Survey volume into a heavily annotated XML version of the volume's text. The input is not raw OCR output but a version in which human OCR correctors have also added pseudo-XML tags to indicate very high level structure corresponding to the nesting of blocks of text. Thus the block of text for a parish contains subordinate blocks for settlements within that parish and they in turn contain subordinate blocks for minor places and street- and field-names located within them. The parish text blocks are themselves contained within blocks for larger historical administrative units such as hundreds, wapentakes, wards, boroughs, chapelries etc. and the block that encompasses all the places within it is the county itself.

In step 1 in Figure 2 the input file is converted from Word (.doc file) to OpenOffice's XML (.odt) format. From this we extract the textual content, the manually added structural tags and all relevant font and style information. The manual structural mark-up is converted to XML elements so that containment relations between places are encoded in the tree-structure of the XML document. From this point all further processing incrementally adds mark-up within the XML structure.

We use the LT-XML2 and LT-TTT2 tools which form the basis of the Edinburgh Geoparser and which have been developed specifically for rule-based processing of text in NLP systems (Grover and Tobin (2006), Tobin et al. (2010)). Along with shell scripting these tools allow us to build up the components that comprise the pipeline. The output of step 2 contains XML elements for paragraphs, sentences and word/punctuation tokens. Font and style information is encoded as attributes on the tokens and line break hyphenation is repaired. Part-of-speech tagging is unnecessary: the named entity classes we recognise are primarily identified by position in the document or on the page in combination with font and style information.

Once the tokens are marked up, finer-grained structural mark-up can be computed inside the wider structure (step 3). Blocks of attestations and etymology descriptions are identified and the title lines of the sections are segmented into elements—e.g. in Figure 1 BARLOW is marked up as the modern name of the place, 97-6428 is recognised as a grid reference and ˈbaːlə is recognised as a pronunciation. Also at this stage, lists of smaller place-, field- and street-names are segmented into individual items.

The NER processing in step 4 uses specially developed rule sets to add detailed mark-up inside sets of attestations. The first line of the attestations in Figure 1 is given the following structure (tokenisation, font and style information suppressed):

```
<altset>
 <alt>
  <histform>Bernlege</histform>
  <attested>
   <date>c. 1030</date>
   <source id="wr796">YCh <item>7</item></source>
  </attested>
 </alt>
</altset>
```

Here YCh is the source of the attestation for the historical form *Bernlege* and the interpretation of the YCh abbreviation is referenced by the id at-

tribute on the source element. Step 4 also uses rule sets to recognise parts of etymological descriptions adding within-sentence mark-up like this:

```
<s> ...
 <etympart>
  <lang>OE</lang>
  <form>bern</form>
  </etympart>
  '<gloss>barn</gloss>' (v.
  <etympart>
   <pn-element>bere-aern</pn-element>
  </etympart>) ...
</s>
```

Step 5 applies rules for non-geographic grounding. For dates, begin and end attributes are computed with obvious values for simple dates and date ranges, a twenty-year window for *circa* dates, other sized windows for century parts (e.g. the first twenty-five years for the early part of a century) and specific periods for regnal dates (Hy 1 denotes Henry I who reigned from 1100 to 1135):

```
<date begin="1086" end="1086">1086</date>
<date begin="1130" end="1139">1130-9</date>
<date begin="1020" end="1040">c. 1030</date>
<date begin="1200" end="1225">e. 13</date>
<date begin="1100" end="1135">Hy 1</date>
```

Place-name elements (<pn-element>) are dealt with at the same stage. These are etymological parts, indicated with bold font in the Survey texts, which belong to a finite set of vocabulary items used in place-names. Place-name elements are catalogued in the Key To English Place-Names (KEPN) database. We look the elements up in KEPN and record their database ID when a match is successful. The final two steps in the processing relate to geographic grounding and are described in more detail in the next section.

We have not been able to perform a formal evaluation of the NER component in the pipeline because we do not have a manually annotated test set. However, we did implement cycles of quality assurance by place-name experts to feed into rule set improvements, so we are confident that the information extracted is of high quality. Our main priority was to capture the historical name attestations for the parishes and main settlements in the Survey. For the blocks that these occur in (e.g. the attestation block in Figure 1) we can get an informal indication of performance by counting the number of non-punctuation tokens that fail to be recognised as part of a historical name or attestation entity. For example, for the three volumes for Derbyshire (published in 1959), there are 342 blocks of main settlement attestations in which our system found 4,052 historical forms associated with 5,817 attestations. There were nine lines of text in these blocks where the processing failed to assign all the words to an entity, resulting in around 20 histform-attestation pairs being missed. Performance is slightly more variable for smaller settlements and lists of streets and field-names, but it is harder to estimate an error rate for these.

3.2 Georeferencing the Survey

To create a historical gazetteer, we need to associate coordinates with every place-name and we do this by aggregating information from several sources and by allowing un-georeferenced place-names to inherit coordinates from a place higher in the XML structure. As described above, some of the Survey volumes associate grid references with some of the places but the coverage is too sporadic to rely on. We therefore use the geographic information in the KEPN database to acquire reliable geo-references as far as possible. KEPN supplies latitude/longitude point coordinates for major settlements (the larger units inside parishes) and we automatically query KEPN for these references, adding the coordinates into the XML in special <geo> elements. For our example of Barlow the <geo> is this:

```
<geo source="kepn" kepnref="14600" long="-1.02337"
     lat="53.7489" placename="Barlow"/>
```

This information is the most authoritative geographic information that we can access but we do not want to discard the other authoritative source of information contained in the grid references in some of the volumes, especially since these may be attached to smaller places not covered by KEPN. We therefore recognise them during the processing, convert them to latitude/longitude and store their coordinates in <geo source="epns"> elements:

```
<geo source="epns" lat="53.7489" long="-1.02337"/>
```

(In this case the coordinates from the two sources are identical but there are cases where they differ slightly.)

Once we have stored KEPN/EPNS geographic information, we implement strategies to achieve high-quality georeferencing of some of the places which do not yet have a georeference. A first step is to utilise the containment relations between places and propagate known georeferences up and down between certain nodes. For example, we do this when a parish with no georeference has the same name as a settlement within it—since they are different administrative levels of the same place, we propagate the <geo> from the settlement to the parish. We aim to provide an authoritative georeference for every parish and

larger settlement in the output, and we have manually built a separate additional resource to supply missing coordinates for 560 parishes/settlements in the entire collection that couldn't be georeferenced using either the volume itself or the KEPN database. While some of these are missing from the database, many are present but couldn't be unambiguously matched because of differences in spelling and punctuation.

At this point there are still many smaller places which do not have a georeference, so we turn to external resources, namely OS data provided through Unlock and the GeoNames gazetteer. We use the geoparser in a non-standard set-up to look up place-names in the external gazetteers and to select the most probable records. To get the results, we feed the geoparser algorithm with the information that we already know from the previous look-up in KEPN/EPNS and we set parameters to choose records which are as close as possible to the known coordinates either of the place-name itself or of its immediate parent or child node. We also apply the geoparser not to whole documents but to individual parishes. This is because the georesolver maximises geographical coherence in its input by choosing coordinates for all the places that will minimize the distance between them—if it is set to work on a single parish, it will automatically tend to select records which are as close to each other as possible. In our running example, Barlow is the first settlement described in the parish of Brayton and the other major settlements within the parish are Brayton, Burn, Gateforth, Hambleton and Thorpe Willoughby. The georesolution algorithm looks at the parish and considers possible groundings of these places together, ensuring as far as possible that the chosen gazetteer records cluster tightly together. If either Unlock or GeoNames does not have a correct entry for one of the places but it does have an entry for somewhere else with the same name, that other entry would be incorrectly chosen. To remedy this situation, we filter the georesolution output and discard any choices which are further than a certain distance (3km) away from coordinates assigned by the earlier KEPN/EPNS step which are known to be correct. This conservative strategy sometimes results in correct groundings being thrown away but it ensures that the Unlock and GeoNames information that we add is highly likely to be correct. The output at this stage for Barlow contains two more <geo> elements in addition to the two already created:

```
<geo source="geonames" gazref="geonames:2656317"
     lat="53.7499300" long="-1.0216400"/>
<geo source="unlock" gazref="unlock:4580690"
     lat="53.74993" long="-1.02164"/>
```

The georeferencing with Unlock and GeoNames considers smaller places as well. In our example there are three minor settlements in the parish, Brayton Barff, Burton Hall and Hambleton Hough, which get assigned coordinates from the OS information in Unlock.

4 User access to the processed data

The XML annotated files that are output from the processing pipeline are an intermediate representation of the information in the Survey which is then converted to other formats in order to make it available to users. Figure 3 summarises how the data is handled after this point. Since the XML output preserves the textual content of the input volumes along with layout, font and style information, it can be used to provide HTML renderings of the text that are visually very similar to the original printed text. The website at EPNS[5] builds on this aspect of the XML output by providing a browse and search interface to all the text associated with a place-name, including both the historical attestation information and the etymological descriptions. Access to this website is restricted to academic users but it provides an invaluable resource for place-name scholars since the information can be searched using the mark-up that we have added (e.g. by map coordinates; by date or source of attestation; by presence of etymological elements or languages, etc.).

On the geographic side, the XML annotated files are converted to a structured format which stores just the historical gazetteer information from the Survey (the DHG). The textual content is discarded while the relevant annotations are transformed into a data structure conforming to the Library of Congress Metadata Authority Description Schema (MADS)[6]. Figure 4 shows parts of the MADS record for Barlow. The modern name is encoded in the mads/authority/geographic element, while the historical variants appear in mads/variant/geographic elements. Geographical coordinates appear in mads/extension/geo elements and attestations, linked to particular variants, are also put in the extension element. The historical forms may be unexpanded shorthands from the original volumes, e.g. *Berlai(a)* meaning either *Berlai* or *Berlaia*, so these are expanded

[5] http://epns.nottingham.ac.uk

[6] http://www.loc.gov/standards/mads/

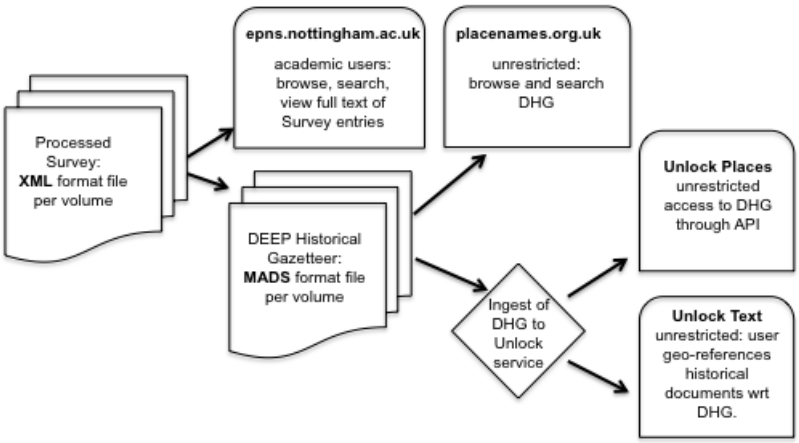


Figure 3: Accessing the DEEP Historical Gazetteer

out in mads/extension/searchterm elements to assist indexing for search. The MADS format feeds all the access mechanisms to the DHG—the data is ingested into the DEEP gazetteer website[7] which allows unrestricted search and browsing access. It is also converted into Unlock's gazetteer format in order that it can be used programmatically via the Unlock API, adding a new resource for users of Unlock Places. The Unlock Text service is one to which users submit documents for geoparsing, and this has been extended to allow them to do this using the DHG. The following section briefly describes how we have adapted the Edinburgh Geoparser for this purpose.

5 Geoparser adaptations

The Edinburgh Geoparser in Unlock standardly georeferences users' documents with reference to the Ordnance Survey and/or GeoNames gazetteers in Unlock. We have reported on this system in Tobin et al. (2010) and evaluated its performance on both modern newspaper text and a variety of historical texts. Other researchers have adapted it for use on different collections of historical text (Isaksen et al., 2011). The two main components of the geoparser are a rule-based NER system for recognising place-names in text and a heuristics-based georesolver to ground the place-names to coordinates (i.e. to choose between competing gazetteer records). In order to update the geoparser to use the historical gazetteer effectively, both of these components need to be extended. We have made the necessary adaptations so that Unlock Text can be used on historical English documents, however it is hard to create a one-size-fits-all version of the system which will perform optimally for all users on all documents—we return to this issue in the final section.

Like many other rule-based NER systems, the NER component in the Edinburgh Geoparser relies in part on lexicons of known entities of relevant types and in part on descriptions of possible contexts for entities encoded as rules. For modern names the geoparser NER system uses extensive place-name lexicons both for Great Britain and globally. To deal with historical names, we converted the MADS-format data into a lexicon of over 500,000 unique entries derived from the searchterms and the modern names and we filtered it to exclude certain lower case forms corresponding to common words. The NER system was given a parameter to specify 'historical mode' and this causes the DHG-derived lexicon to be applied instead of the modern place-name lexicons. Rules for place-name contexts apply as usual, as do rules and lexical look-up for other entity types.

For the georesolution component, the DHG was added to the list of available gazetteers. Using it results in a set of records with associated coordinates that need to be disambiguated in order to ground the place-names. This is sufficient for use of the DHG in georeferencing but there are some extra functionalities that suggest themselves in this context. The first concerns the users' knowledge about the geographic focus of their documents: assuming they know that the document is about a particular county or sub-area of England, it is useful to constrain the georeferencing results to exclude out-of-area interpretations. To achieve this we allow the user to specify one or more of the DHG counties as a constraint. A second extension follows from the fact that Unlock returns DHG records that include date of attestation. We have

[7] http://placenames.org.uk

```
<mads ID="epns-deep-33-b-subparish-000011">
  <authority ID="33-b-name-subparish-000011">
   <geographic valueURI="http://placenames.org.uk/id/placename/33/001099">Barlow</geographic>
  </authority>
  <related type="broader" xlink:href="#33-a-parish-000004">
   <geographic>Brayton</geographic>
  </related>
  <variant ID="33-b-name-w52628">
   <geographic valueURI="http://placenames.org.uk/id/placename/33/001100">Bernlege</geographic>
  </variant>
  <variant ID="33-b-name-w52652">
   <geographic valueURI="http://placenames.org.uk/id/placename/33/001101">
     Berlai(a), Berley(e), Berlay(e)</geographic>
  </variant>
......
  <recordInfo>
   <recordCreationDate>2013-10-10</recordCreationDate>
   <recordContentSource valueURI=
"http://epns.nottingham.ac.uk/England/West%20Riding%20of%20Yorkshire/Barkston%20Ash%20Wapentake/Brayton/Barlow"/>
  </recordInfo>
  <extension>
   <geo source="geonames" gazref="geonames:2656317" lat="53.7499300" long="-1.0216400"/>
   <geo source="epns" lat="53.74422247" long="-1.029470762"/>
   <geo source="unlock" gazref="unlock:11070229" lat="53.74865601989839" long="-1.021785033055991"/>
   <geo source="kepn" kepnref="14600" lat="53.7489" long="-1.02337"/>
   <attestation variantID="33-b-name-w52628">
    <date subtype="circa" begin="1020" end="1040">c. 1030</date>
    <source id="wr796" style="">YCh</source>
    <item>7</item>
   </attestation>
   <attestation variantID="33-b-name-w52652">
    <date subtype="simple" begin="1086" end="1086">1086</date>
    <source id="wr123" style="">DB</source>
   </attestation>
......
   <searchterm variantID="33-b-name-w52628">Bernlege</searchterm>
   <searchterm variantID="33-b-name-w52652">Berlaia</searchterm>
   <searchterm variantID="33-b-name-w52652">Berlai</searchterm>
......
  </extension>
 </mads>
```

Figure 4: MADS Sample (redacted)

adapted the geoparser to allow the user to specify a date range as a constraint.

Figure 5 shows a screenshot of our development visualisation tool where we have used the adapted geoparser to georeference a Dorset Feet of Fines document from the Internet Archive[8]. The geoparser was run with the county constraint set to 'Dorset' in order to exclude any possible matches from outwith that county. The display shows a map with purple (darker) pins for the preferred groundings of the places that were recognised and could be grounded, and green (lighter) pins for alternative possible groundings. The scrollable text pane shows the text with place-name entities highlighted (ones which are links are those that have been successfully grounded). The third pane shows the coordinates for the gazetteer entries that have been returned. The first (purple) coordinates are the preferred ones and the remaining (green) ones are lower ranked alternatives. Note that because we use only the historical gazetteer and a Dorset constraint, several of the modern names are not grounded (e.g. *Westminster*, *Taunton*). The correct *Westminster* is in the Survey under Middlesex and therefore not accessed. In the case of *Taunton*, there are two instances in the DHG: a modern name for a minor settlement in Surrey and a historical form of modern *Taynton* in Gloucestershire. The actual interpretation is likely to be modern Taunton in Somerset, which is one of the counties not yet in the Survey. Several of the historical place-names recognised in the text have not been grounded to a place in Dorset. There are entries in the DHG for some of these, e.g. a *Bundebi* in Lincolnshire and a *Rading* in Berkshire.

6 Discussion

The example in Figure 5 is intended to illustrate some of the issues that are involved in using the geoparser on a particular historical text. The user who wants this particular text georeferenced has a number of options. Without using any constraint on the area to be considered, many of the place-names would be wrongly grounded. The Dorset-only constraint is probably too conservative and the user might instead prefer to use a different option available as standard with the geoparser which is to specify a bounding box or circle to weight the entries within them more highly. This option differs from the Dorset-only constraint, which only considers DHG entries known to be in Dorset, in

[8]This is the full text, i.e. OCR-ed version, of the document at `https://archive.org/details/fullabstractsfe00frygoog`.

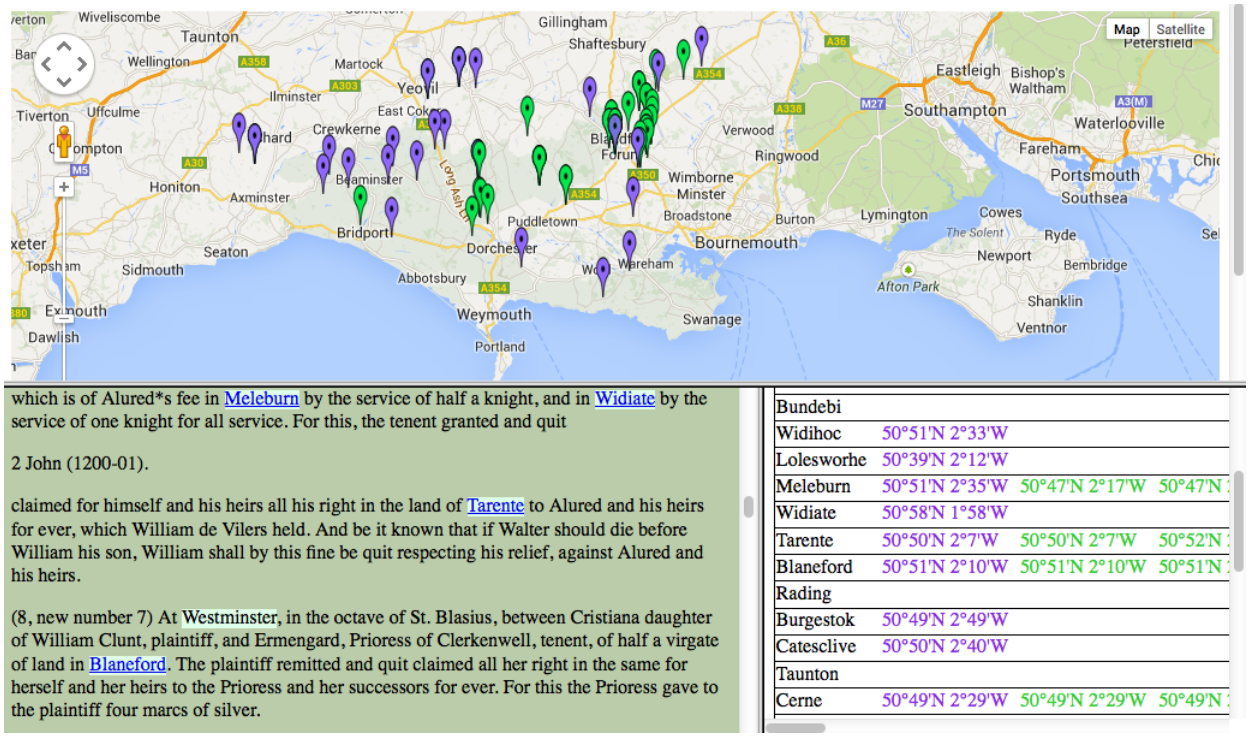


Figure 5: Visualisation of geoparser output on Dorset Feet of Fines

that it considers entries from all of the counties and influences the rankings of the possible entries. The user can follow it with a clean-up stage to remove groundings which fall outside the bounding box or circle. To get interpretations for place-names like *Westminster* and *Taunton*, the user could submit the document to a second run of the geoparser using the modern OS gazetteer and then combine the results of the two runs. Alternatively, the user might opt to manually post-edit the output of the geoparser: a tool would be useful for this so we are planning to add a map-based georeference annotation capability to the geoparser.

The Edinburgh Geoparser is available as a service from Unlock Text but there are so many types of historical document and so many user needs, that it is unlikely to provide all the possible options and flexibility that might be required. For this reason we anticipate that many users will prefer to access the DHG via Unlock Places for integration with their own systems; other users will want access to the source of the geoparser in order to tailor it for their specific needs. An open source version will shortly be available from `http://www.ltg.ed.ac.uk`.

The data described here is not linked data in the usual sense of the term (i.e. it is not RDF). However, we have been careful to add as many linkages as we can. The core data structure is the MADS data collection (Figure 4) and this contains two kinds of URI: in the valueURI attribute on mads/authority/geographic there is a link to the relevant page on the EPNS website, while in the valueURI attribute on mads/recordInfo/recordContentSource there is a link to the placenames.org.uk site. Three of the mads/extension/geo elements contain references to external data sources: the kepnref id points to the KEPN database and gazref ids point to the relevant records in Unlock and GeoNames. Because the MADS data collection conforms to a recognised standard, it would be relatively easy to convert it to RDF and publish it as linked data. Moreover, the Unlock version of the DHG retains all the information in the MADS collection and this means that the output of the geoparser can be made to retain the links out from the entries, enabling the user to link their historical texts to the DHG and to KEPN, Unlock and GeoNames.

Acknowledgements

DEEP was funded by JISC's Digitisation and Content Programme. We are indebted to our partners from the Institute for Name-Studies at the University of Nottingham, the Centre for Data Digitization and Analysis at Queen's University Belfast, the Centre for eResearch at King's College London and EDINA at the University of Edinburgh.

References

Gregory Crane. 2004. Georeferencing in historical collections. *D-Lib Magazine*, 10(5).

Claire Grover and Richard Tobin. 2006. Rule-based chunking and reusability. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society A*.

Uta Hinrichs, Beatrice Alex, Jim Clifford, and Aaron Quigley. to appear 2014. Trading consequences: A case study of combining text mining and visualisation to facilitate document exploration. In *Digital Humanities 2014*.

Tim Hitchcock, Robert Shoemaker, and Jane Winters. 2011. Connected Histories: A new web search tool for British historians. *History*, 96(323):354–356.

Leif Isaksen, Elton Barker, Eric C. Kansa, and Kate Byrne. 2011. GAP: A NeoGeo Approach to Classical Resources. *Leonardo Transactions*, 45(1).

C.J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, and Patricia Murrieta-Flores. 2013. Customising geoparsing and georeferencing for historical texts. In *2013 IEEE International Conference on Big Data*, pages 59–62.

Richard Tobin, Claire Grover, Kate Byrne, James Reid, and Jo Walsh. 2010. Evaluation of georeferencing. In *Proceedings of Workshop on Geographic Information Retrieval (GIR'10)*.